Integrated Machine Learning and CNN Approaches for Breast Cancer Prediction Using Mammography Images

Rajendra Randa[1], Sanjeev Gour[2]

[1]*Research Scholar, Dept. Of Computer Science at Medi-Caps University Indore, India,*

[2]*Research Supervisor and Asst. Professor, Dept. Of Computer Science at Medi-Caps University Indore, India*

KEYWORDS

Machine learning

CNN

Mammography Images

Predictive Healthcare

ABSTRACT:

Breast Cancer has increased in the last few years not only in females but also in males, but majorly it affects women's lives. At the rate at which this issue occurs, we need advanced tools, treatments, and methods to predict breast cancer. Dl and ML are essential parts of AI technology that can be helpful in various domains such as finance, banking, cyber security, and healthcare. With emerging technology, we have various types of data from healthcare settings to test the diseases more robustly, and AI and their related technology are very helpful in analyzing the data and finding out outcomes from patient records to finalize better treatment for it in early stages and to provide a healthy life. DL and ML are widely used for classification problems on imaging data. In this paper, we studied various machine learning models and CNN to find out which model is compatible with medical images to classify them and provide better outcomes. This paper used LR, NB, KNN, Support vector machine, GB, Xtreme gradient boosting, and CNN methods to classify images for malignant and benign images. As for future work, we can study various advanced ensemble learning and deep learning approaches. So that we can predict various types of sensitive medical data more accurately.

1. Introduction

Machine Learning and deep learning are essential to artificial intelligence. Supervised learning and CNN algorithms are widely used for image classifications in the healthcare sector. In healthcare, machine learning and deep learning classify imaging data and predict accurate outcomes using historical training datasets. In cancer disease prediction machine learning and other related technologies are useful to predict cancer more robustly for clinical and imaging medical data. Data can be x-rays, mammography, MRIs, etc used as medical imaging data. Machine learning algorithms based on linear structure and deep learning algorithms work in the form of neurons. In this study, we are using traditional machine learning approaches and CNN on imaging datasets to see their compatibility. Machine learning and deep learning can classify image data very robustly in healthcare settings.

.

2. Literature Review

Machine learning is used to solve classification problems. When we get unstructured data or non-textual datasets, such as mammography data, in healthcare, we also use advanced machine learning models and other deep learning models.

Recent advancements in deep learning have significantly enhanced breast cancer detection and diagnosis through various imaging modalities and analytical techniques. For instance, Reshma et al. (2022) utilized deep learning methods on histopathological images, achieving notable improvements in classification accuracy. Similarly, Lehman et al. (2017) demonstrated that deep learning algorithms could match or even surpass radiologists in diagnostic accuracy for mammography, highlighting the potential of AI in clinical settings. In another study, McKinney et al. (2020) developed an AI system that outperformed human radiologists in breast cancer detection on mammograms, reducing both false positives and false negatives. This system was trained on a large dataset from the UK and the USA, showcasing its robustness across diverse populations. Sadhukhan et al. (2019) explored the integration of image processing and machine learning techniques for breast cancer diagnosis, emphasizing the importance of feature extraction in improving model performance. Yousefi et al. (2019) investigated the application of deep learning in genomics, providing insights into how these methods can be applied to genetic data for cancer diagnosis. Additionally, Jabeen et al. (2020) conducted a comprehensive review of deep learning-based computer-aided diagnosis systems for breast cancer, discussing various architectures and their effectiveness in different diagnostic scenarios. In the realm of wearable technology, Khan et al. (2020) proposed a smart wearable device for early detection of breast cancer, utilizing machine learning algorithms to analyze physiological data. Zhang et al. (2022) examined the role of deep learning in breast cancer histopathological image analysis, highlighting the advancements in automated feature extraction and classification. Moreover, Wang et al. (2022) discussed the integration of deep learning with radiomics, further improving breast cancer detection by analyzing mammographic images. Collectively, these studies underscore the transformative potential of deep learning and machine learning techniques in enhancing breast cancer detection and diagnosis across various modalities, including histopathological images, mammograms, genomics, and wearable technology. Deep learning is widely used to conduct a robust study on medical imaging data; sometimes machine learning approaches are also helpful.

.

3. Methods

To study we follow step by step approach to find our expected outcomes and try to fill the gap in the studied literature. These are various steps taken in the study of breast cancer prediction for medical image data.

Dataset: This dataset is taken from a public data repository (source: https://www.kaggle.com/datasets/abdulrazakadekunle/mammogram-mastery-a-robust-dataset-breast-cancer?resource=download). The dataset has 745 images with having two categories of images with labelled ‘cancer’ and ’non-cancer’ images. To make predictions we separate the data into 80-20 forms to use in training and testing. Here malignant" stands for cancerous or cancer positive, while "benign" means non-cancerous or considered as cancer negative; so, a malignant tumor is a cancerous growth, and a benign tumor is not cancerous.

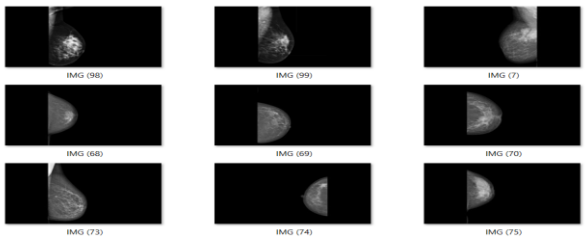


Fig 1. sample of used imaging dataset used in an experiment.

Data Preprocessing and Analysis: we have imaging data with two labels malignant as 'cancer' and benign as 'non-cancer. Before splitting the data into training and testing sets, we resize the all images into the same format and size to get better outcomes. After that, we put labels (0 and 1) where 1 represents cancer and 0 stands for non-cancer images. After classifying images into label-wise categories, we transform them into a structured format.

Machine Learning: Machine learning is the set of algorithms to solve predictive problems with better accuracy and faster than traditional methods from complex structured and unstructured data in various fields. In this paper, we used LR, NB, KNN, Support vector machine, GB and Xtreme gradient boosting algorithms.

- **LR:** logistic regression works well on text data and also with imaging data to classify results. Logistic regression provides better compatibility with binary classification problems.
- **NB:** it is a prior and posterior probability-based model to classify the text and imaging data.
- **KNN:** it is a cluster-based algorithm based on the total number of nearest elements from the target element.
- **Support Vector Machine:** SVM builds a best-fit plan between different classes of data with maximum margin based on the number of support vector data points.
- **GB:** it is an ensemble learning technique to improve the accuracy of classification and also regression problems and is based on boosting approaches on functional space where the target is treated as pseudo-residuals.
- **XG-boosting:** It is also an implementation of gradient boosting ensemble method to get higher accuracy results from complex and grid types of data.

Deep Learning (CNN): Deep learning is an essential part of machine learning; deep learning models work based on neurons with activation functions. Machine learning also works on classification problems on images as well and deep learning is most impactful for it. In our study, we used some traditional ML and CNN algorithms to make the early-stage breast cancer using mammography images, and then find compatible algorithms from them based on their accuracy.

Predictive analysis: we applied serval models to study their compatibilities on grid or imaging data to classify cancer and help the medical professional to improve the quality of treatment and increase the speed to prioritise the personalized treatments based on the patient's history.

These are the steps following to experiment successfully to predict breast cancer using imaging data classification.

4. Experimental Setup

For the experimental set we used the i5 12th gen system for better performance and an anaconda navigator as IDLE to perform the task. Then to predict breast cancer we find the dataset as per our requirements from the medical centre or an online resource and then perform the machine learning and deep learning tasks on available imaging data. To apply ml and dl tasks, we use the following frameworks:

Data analysis and preprocessing: to import and analyze the data, we use NumPy, pandas, and os frameworks, and then we import the dataset and then reshape the dataset and provide them labels such as 'cancer' and 'non-cancer' to predict the outcome effectively. After analyzing the dataset, we split the dataset into training and testing (80% and 20%) to use the training dataset as the primary dataset and after building the model, we used the testing dataset to validate the performance of existing models.

Data visualization: we visualise the result, and comparison chart in graphical mode to make it understand easily and for it, we are using matplotlib and seaborn libraries or frameworks. To purpose of visualising the features is to improve the analytical skills during the experiment and it is the better way to analyse and compare the data from results, comparison of accuracies and colouration of features during the analysis of data.

Machine learning: to apply machine learning models, we import 'sci-kit learn' and 'xgboost' libraries with their essential features such as 'linear_model' for logistics, 'neighbour' for KNC, 'ensemble' for GB and RF, naïve_baiys for NB, SVM for support vector classifiers and xgboost for XGBClassifire. These are the models we used in our study to predict breast cancer effectively.

Deep learning: In our study, we are using the traditional CNN approach to predict we import 'tensorflow' and 'keras' to implement the CNN model here we use custom layer traditional CNN in the place of pre-trained CNN approaches to identify how CNN works on imaging data on the comparison of machine learning approaches.

Model's Performance: to validate the performance we visualized ROC-Curve and bar chart using matplotlib and seaborn libraries. And split the dataset using a linear model to get the rest of the unseen data as testing data to validate the performances.

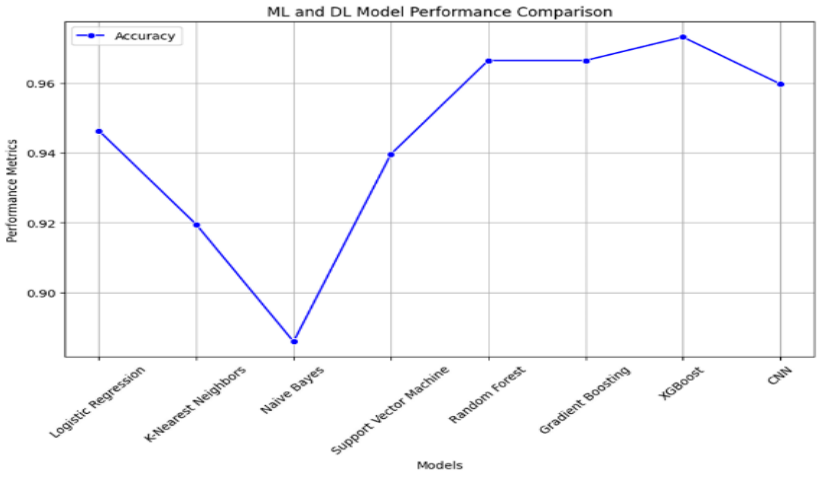
5. Results

breast cancer is leading the attention of medical professional because of its mortality rates in worldwide areas. DL, ML and Ensemble methods are used to classification problems to predict diseases using text and imaging data. This study provides comparative and practical analysis for imaging breast cancer data that contain low- and high-quality images. Using AI technology helps to improve accuracy and reduce anomalies.

Machine learning can also predict diseases using grid data but some model's accuracy and compatibilities affect the quality of data, but ensemble learning and deep learning methods work better with high and low-quality data during training and testing data. But with high-quality imaging data machine learning algorithms such as LG, RF, NB, etc also work with better accuracies. The following line chart is a graphical representation of various accuracies of models used in experiments. The X-axis contains model names and y contains accuracies level from 0-1. In our study, we consider CNN

from deep learning, as SL we used logistic, random forest, support vector machine, and naïve bays and as ensemble learning, we used gradient boosting and extreme gradient boosting.

Fig 2. ML and DL performance on breast cancer imaging data



The following table contains the details of each model used in our study with their accuracies and other evolution values such as f1-score. Recall and precision. And as per our study, we used traditional machine learning algorithms and CNN models.

Model	Accuracy	F1-score	Recall	Precision
LR	0.946309	0.750000	0.60	1.000000
KNN	0.919463	0.571429	0.40	1.000000
NB	0.885906	0.585366	0.60	0.571429
SVM	0.939597	0.709677	0.55	1.000000
RF	0.966443	0.857143	0.75	1.000000
GB	0.966443	0.857143	0.75	1.000000
XGB	0.973154	0.888889	0.80	1.000000
CNN	0.959732	NaN	NaN	NaN

Table 1. Comparison of various algorithms

During training overall all selected models provide similar and better accuracies if we use high-quality images but during low-quality images, NB provide similar compatibilities during training and testing data but logistics (94%), and KNN (91%) work well with good quality data but with decreasing the quality of images differences between their accuracies during training and testing are going to increase. SVM provide similar accuracies (93%) with high and low-quality images, and RF provide better accuracy (96%) with images but with very low-quality images it also provides good accuracy

but ensemble techniques such as GB and XGB perform very well with all types of imaging data with similar compatibilities. And CNN performs robustly with low and high-quality medical-images.

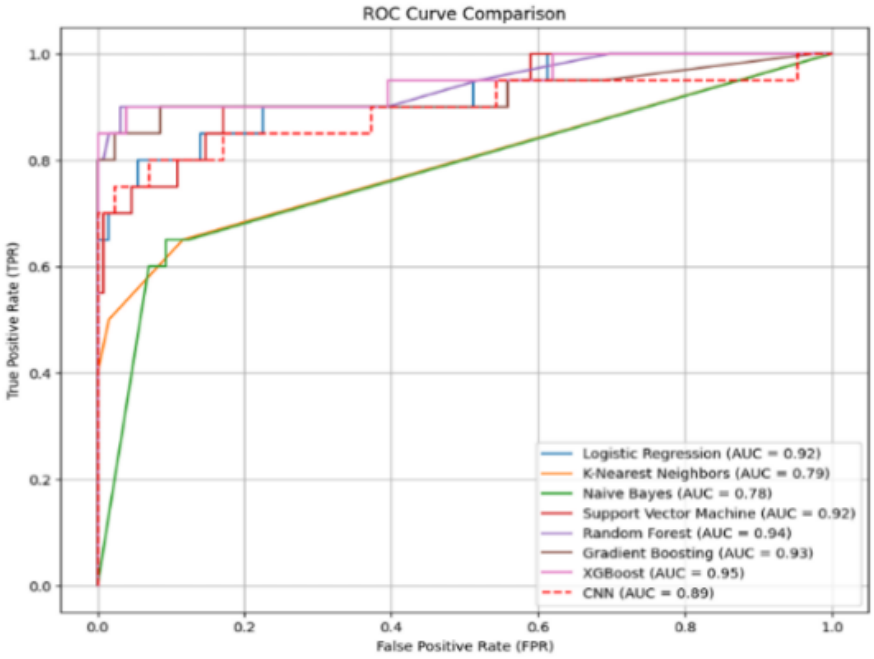


Fig 3. ROC curve comparison

The above graph represents the study of the evolution ROC curve of our study. It is a graphical representation to show how well any model performs during binary classification at different threshold values and it plots between TPR and FPR. It is the best way to summarize the outcomes of machine learning and ensemble learning technology.

6. Discussion

Breast Cancer is one of the leading issues in various types of cancer and also in women's diseases. In the area of cancer treatment, we have two types of data to predict the diseases one is text clinical data and another one is imaging data, one provides details about the medical conditions of patients and the other type of data represents an actual condition of cells and organs. Using advanced methods in technology improves the quality of treatment and survival rate in real-world healthcare settings. This study tried to fill the gap to identify the compatibility of integrated ML models and CNN on high and low-quality imaging data and how they affect outcomes. This practical approach defines that integrated and advanced ml algorithms and CNN perform better than traditional approached on low-quality images and as reference work we can try to apply ensemble techniques on machine learning and deep learning algorithms to improve the quality of outcomes and increase the capabilities of learning models on healthcare domain.

7. Conclusion

This paper provides a comprehensive analysis of mammographic images to predict cancer using CNN and traditional machine learning models. We predict cancer with two binary classifications of data such as malignant and benign. Malignant covers the patients who have cancer and benign for the non-cancer patients. In this paper, we studied a few ML algorithms such as LR, SVM and others and the CNN model to classify imaging data to predict cancer. Machine Learning and deep learning both can classify medical imaging data but when it comes to the size and celerity of images CNN responds better than some machine learning models. Some ML algorithms such as XGB, gradient boosting and RF perform better during training and testing duration.

In future work, we should try to apply advanced ML, EL, DL and GPTs to make predictions effectively. Using advanced models, we can predict diseases with high- and low-quality imaging data technology is growing day by day and we will try to use advanced AI tools and models to predict the diseases in early stage in real-world patient data to improve the survival chances and decrease the mortality rates.

Refrences

1. Prasad Jasti, V. D., Zamani, A. S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., Raghuvanshi, A., & Kaliyaperumal, K. (2021). Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. Security and Communication Networks, 2022(1), 1918379. https://doi.org/10.1155/2022/1918379
2. Couture, H. D., Williams, L. A., Geradts, J., Nyante, S. J., Butler, E. N., Marron, J. S., Perou, C. M., Troester, M. A., & Niethammer, M. (2018). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. Npj Breast Cancer, 4(1), 1-8. https://doi.org/10.1038/s41523-018-0079-1
3. Couture, H. D., Williams, L. A., Geradts, J., Nyante, S. J., Butler, E. N., Marron, J. S., Perou, C. M., Troester, M. A., & Niethammer, M. (2018). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. Npj Breast Cancer, 4(1), 1-8. https://doi.org/10.1038/s41523-018-0079-1
4. Xie, J., Liu, R., Luttrell, J., & Zhang, C. (2019). Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Frontiers in Genetics, 10, 426920. https://doi.org/10.3389/fgene.2019.00080
5. Zerouaoui, H., Idri, A. Reviewing Machine Learning and Image Processing Based Decision-Making Systems for Breast Cancer Imaging. J Med Syst 45, 8 (2021). https://doi.org/10.1007/s10916-020-01689-1
6. Rakhlin, A., Shvets, A., Iglovikov, V., Kalinin, A.A. (2018). Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. In: Campilho, A., Karray, F., ter Haar Romeny, B. (eds) Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science, vol 10882. Springer, Cham. https://doi.org/10.1007/978-3-319-93000-8_83
7. Jabeen, S., Khan, M. A., Alhaisoni, M., Tariq, U., Zhang, Y. D., & Hamza, A. (2020). Computer-aided diagnosis for breast cancer classification using deep learning: A comparative approach. *Computers, Materials & Continua, 63*(3), 1241–1263. https://doi.org/10.32604/cmc.2020.011909
8. Khan, M. A., Ashraf, I., Alhaisoni, M., Tariq, U., Damaševičius, R., Scherer, R., & Kadry, S. (2020). A novel deep learning-based framework for an intelligent wearable system to monitor

patients with breast cancer. *IEEE Access, 8*, 135776–135785. https://doi.org/10.1109/ACCESS.2020.3011627

9. Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2017). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine, 177*(9), 1277–1283. https://doi.org/10.1001/jamainternmed.2017.2764
10. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Suleiman, O. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*, 89–94. https://doi.org/10.1038/s41586-019-1799-6
11. Reshma, P. R., Sundararajan, M., & Kumar, A. S. (2022). Histopathological image classification for breast cancer detection using deep learning techniques. *Journal of Healthcare Engineering, 2022*, 1–10. https://doi.org/10.1155/2022/2895647
12. Sadhukhan, S., Bhattacharjee, D., & Nasipuri, M. (2019). An approach for automatic mass detection in mammography images using image processing and machine learning techniques. *Biomedical Signal Processing and Control, 52*, 72–82. https://doi.org/10.1016/j.bspc.2019.03.006
13. Wang, J., Zhang, X., & Yu, S. (2022). Deep learning and radiomics in breast cancer diagnosis: Recent advances and future trends. *Frontiers in Oncology, 12*, 896234. https://doi.org/10.3389/fonc.2022.896234
14. Yousefi, S., Luo, X., Sun, Y., Elloumi, F., Berman, D., Zanetti, C., Fan, J. B., & Chen, X. (2019). Deep learning-based genomics analysis for precision oncology. *npj Precision Oncology, 3*(1), 1–9. https://doi.org/10.1038/s41698-019-0087-6
15. Zhang, L., Wang, Y., Yang, Y., & Liu, H. (2022). A deep learning approach for histopathological image analysis in breast cancer diagnosis. *Computers in Biology and Medicine, 141*, 105018. https://doi.org/10.1016/j.compbiomed.2021.105018
16. Koshy, S.S., Anbarasi, L.J., Jawahar, M. et al. Breast cancer image analysis using deep learning techniques – a survey. Health Technol. 12, 1133–1155 (2022). https://doi.org/10.1007/s12553-022-00703-5
17. Furkan Atban, Ekin Ekinci, Zeynep Garip, Traditional machine learning algorithms for breast cancer image classification with optimized deep features, Biomedical Signal Processing and Control,Volume 81,2023,104534,ISSN 1746-8094,https://doi.org/10.1016/j.bspc.2022.104534.(https://www.sciencedirect.com/science/article/pii/S1746809422009880)
18. Reshma VK, Arya N, Ahmad SS, Wattar I, Mekala S, Joshi S, Krah D. Detection of Breast Cancer Using Histopathological Image Classification Dataset with Deep Learning Techniques. Biomed Res Int. 2022 Mar 3;2022:8363850. doi: 10.1155/2022/8363850. Retraction in: Biomed Res Int. 2024 Mar 20;2024:9863139. doi: 10.1155/2024/9863139. PMID: 35281604; PMCID: PMC8913119.
19. Hameed, Z., Zahia, S., Javier Aguirre, J., & María Vanegas, A. (2019). Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models. Sensors, 20(16), 4373. https://doi.org/10.3390/s20164373

20. Boumaraf, S., Liu, X., Wan, Y., Zheng, Z., Ferkous, C., Ma, X., Li, Z., & Bardou, D. (2021). Conventional Machine Learning versus Deep Learning for Magnification Dependent Histopathological Breast Cancer Image Classification: A Comparative Study with Visual Explanation. Diagnostics, 11(3), 528. https://doi.org/10.3390/diagnostics11030528
21. Chaudhury, S., Krishna, A. N., Gupta, S., Sankaran, K. S., Khan, S., Sau, K., Raghuvanshi, A., & Sammy, F. (2021). [Retracted] Effective Image Processing and Segmentation-Based Machine Learning Techniques for Diagnosis of Breast Cancer. Computational and Mathematical Methods in Medicine, 2022(1), 6841334. https://doi.org/10.1155/2022/6841334
22. Sadhukhan, S., Upadhyay, N., Chakraborty, P. (2020). Breast Cancer Diagnosis Using Image Processing and Machine Learning. In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_12
23. Prasad Jasti, V. D., Zamani, A. S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., Raghuvanshi, A., & Kaliyaperumal, K. (2021). Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. Security and Communication Networks, 2022(1), 1918379. https://doi.org/10.1155/2022/1918379
24. Luuk Balkenende, Jonas Teuwen, Ritse M. Mann,Application of Deep Learning in Breast Cancer Imaging,Seminars in Nuclear Medicine,Volume 52, Issue 5, 2022, Pages 584-596, ISSN 0001-2998, https://doi.org/10.1053/j.semnuclmed.2022.02.003.
25. Nahid, Abdullah-Al, Kong, Yinan, Involvement of Machine Learning for Breast Cancer Image Classification: A Survey, Computational and Mathematical Methods in Medicine, 2017, 3781951, 29 pages, 2017. https://doi.org/10.1155/2017/3781951.
26. Thakur, R., & Panse, P. (2023). Machine learning techniques applied for land use—land cover (LULC) image classification: Research avenues challenges with issues. In Mobile radio communications and 5G networks: Proceedings of Third MRCN 2022 (pp. 281–296). Springer Nature Singapore.
27. Thakur, R., & Panse, P. (2022). Classification performance of land use from multispectral remote sensing images using a decision tree, K-nearest neighbour, random forest, and support vector machine using EuroSAT data. *International Journal of Intelligent Systems and Applications in Engineering, 10*(1), 67–77.
28. Gour, S., Qureshi, A. R. K., Tukra, G. S., Bairagi, V., Dehariya, H., & Shandilya, A. (2024). Impact of machine learning techniques in medical treatment assistance in perspective to AR & VR technologies. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202449103006